

Appendix A. Proofs of Theorems in the Paper

A.1. Proposition and lemma

The following proposition holds for the mini-batch gradient.

Proposition A.1. *Let $t \in \mathbb{N}$, ξ_t be a random variable independent of ξ_j ($j \in [0 : t-1]$), $\theta_t \in \mathbb{R}^d$ be independent of ξ_t , and $\nabla f_{B_t}(\theta_t)$ be the mini-batch gradient, where $f_{\xi_{t,i}}$ ($i \in [b_t]$) is the stochastic gradient (see Assumption 1(A2)). Then, the following hold:*

$$\mathbb{E}_{\xi_t} \left[\nabla f_{B_t}(\theta_t) \middle| \hat{\xi}_{t-1} \right] = \nabla f(\theta_t) \text{ and } \mathbb{V}_{\xi_t} \left[\nabla f_{B_t}(\theta_t) \middle| \hat{\xi}_{t-1} \right] \leq \frac{\sigma^2}{b_t},$$

where $\mathbb{E}_{\xi_t}[\cdot | \hat{\xi}_{t-1}]$ and $\mathbb{V}_{\xi_t}[\cdot | \hat{\xi}_{t-1}]$ are respectively the expectation and variance with respect to ξ_t conditioned on $\xi_{t-1} = \hat{\xi}_{t-1}$.

Proof Assumption 1(A3) and the independence of b_t and ξ_t ensure that

$$\mathbb{E}_{\xi_t} \left[\nabla f_{B_t}(\theta_t) \middle| \hat{\xi}_{t-1} \right] = \mathbb{E}_{\xi_t} \left[\frac{1}{b_t} \sum_{i=1}^{b_t} \nabla f_{\xi_{t,i}}(\theta_t) \middle| \hat{\xi}_{t-1} \right] = \frac{1}{b_t} \sum_{i=1}^{b_t} \mathbb{E}_{\xi_{t,i}} \left[\nabla f_{\xi_{t,i}}(\theta_t) \middle| \hat{\xi}_{t-1} \right],$$

which, together with Assumption 1(A2)(i) and the independence of ξ_t and ξ_{t-1} , implies that

$$\mathbb{E}_{\xi_t} \left[\nabla f_{B_t}(\theta_t) \middle| \hat{\xi}_{t-1} \right] = \frac{1}{b_t} \sum_{i=1}^{b_t} \nabla f(\theta_t) = \nabla f(\theta_t). \quad (6)$$

Assumption 1(A3), the independence of b_t and ξ_t , and (6) imply that

$$\begin{aligned} \mathbb{V}_{\xi_t} \left[\nabla f_{B_t}(\theta_t) \middle| \hat{\xi}_{t-1} \right] &= \mathbb{E}_{\xi_t} \left[\left\| \nabla f_{B_t}(\theta_t) - \nabla f(\theta_t) \right\|^2 \middle| \hat{\xi}_{t-1} \right] \\ &= \mathbb{E}_{\xi_t} \left[\left\| \frac{1}{b_t} \sum_{i=1}^{b_t} \nabla f_{\xi_{t,i}}(\theta_t) - \nabla f(\theta_t) \right\|^2 \middle| \hat{\xi}_{t-1} \right] \\ &= \frac{1}{b_t^2} \mathbb{E}_{\xi_t} \left[\left\| \sum_{i=1}^{b_t} (\nabla f_{\xi_{t,i}}(\theta_t) - \nabla f(\theta_t)) \right\|^2 \middle| \hat{\xi}_{t-1} \right]. \end{aligned}$$

From the independence of $\xi_{t,i}$ and $\xi_{t,j}$ ($i \neq j$) and Assumption 1(A2)(i), for all $i, j \in [b_t]$ such that $i \neq j$,

$$\begin{aligned} &\mathbb{E}_{\xi_{t,i}} [\langle \nabla f_{\xi_{t,i}}(\theta_t) - \nabla f(\theta_t), \nabla f_{\xi_{t,j}}(\theta_t) - \nabla f(\theta_t) \rangle | \hat{\xi}_{t-1}] \\ &= \langle \mathbb{E}_{\xi_{t,i}} [\nabla f_{\xi_{t,i}}(\theta_t) | \hat{\xi}_{t-1}] - \mathbb{E}_{\xi_{t,i}} [\nabla f(\theta_t) | \hat{\xi}_{t-1}], \nabla f_{\xi_{t,j}}(\theta_t) - \nabla f(\theta_t) \rangle \\ &= 0. \end{aligned}$$

Hence, Assumption 1(A2)(ii) guarantees that

$$\mathbb{V}_{\xi_t} \left[\nabla f_{B_t}(\theta_t) \middle| \hat{\xi}_{t-1} \right] = \frac{1}{b_t^2} \sum_{i=1}^{b_t} \mathbb{E}_{\xi_{t,i}} \left[\left\| \nabla f_{\xi_{t,i}}(\theta_t) - \nabla f(\theta_t) \right\|^2 \middle| \hat{\xi}_{t-1} \right] \leq \frac{\sigma^2 b_t}{b_t^2} = \frac{\sigma^2}{b_t},$$

which completes the proof. \square

Motivated by Lemma 1 in (Liu et al., 2020), we prove the following lemma.

Lemma A.1. *Under Assumption 1, Algorithm 1 satisfies that, for all $t \in \{0\} \cup \mathbb{N}$,*

$$\mathbb{E} \left[\left\| \mathbf{m}_t - (1 - \beta) \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) \right\|^2 \right] \leq (1 - \beta)^2 \sigma^2 \sum_{i=0}^t \frac{\beta^{2(t-i)}}{b_i},$$

where \mathbb{E} denotes the total expectation defined by $\mathbb{E} := \mathbb{E}_{\boldsymbol{\xi}_0} \mathbb{E}_{\boldsymbol{\xi}_1} \cdots \mathbb{E}_{\boldsymbol{\xi}_t}$.

Proof The definition of \mathbf{m}_t and $\mathbf{m}_{-1} := \mathbf{0}$ ensure that

$$\begin{aligned} \mathbf{m}_t &= \beta \mathbf{m}_{t-1} + (1 - \beta) \nabla f_{B_t}(\boldsymbol{\theta}_t) \\ &= \beta \{ \beta \mathbf{m}_{t-2} + (1 - \beta) \nabla f_{B_{t-1}}(\boldsymbol{\theta}_{t-1}) \} + (1 - \beta) \nabla f_{B_t}(\boldsymbol{\theta}_t) \\ &= \beta^2 \mathbf{m}_{t-2} + (1 - \beta) \{ \beta \nabla f_{B_{t-1}}(\boldsymbol{\theta}_{t-1}) + \beta^0 \nabla f_{B_t}(\boldsymbol{\theta}_t) \} \\ &= \beta^{t+1} \mathbf{m}_{-1} + (1 - \beta) \sum_{i=0}^t \beta^{t-i} \nabla f_{B_i}(\boldsymbol{\theta}_i) \\ &= (1 - \beta) \sum_{i=0}^t \beta^{t-i} \nabla f_{B_i}(\boldsymbol{\theta}_i), \end{aligned}$$

which, together with $\|\boldsymbol{\theta}\|^2 = \langle \boldsymbol{\theta}, \boldsymbol{\theta} \rangle$, implies that

$$\begin{aligned} &\left\| \mathbf{m}_t - (1 - \beta) \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) \right\|^2 \\ &= (1 - \beta)^2 \left\| \sum_{i=0}^t \beta^{t-i} (\nabla f_{B_i}(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_i)) \right\|^2 \\ &= (1 - \beta)^2 \sum_{i=0}^t \sum_{j=0}^t \langle \beta^{t-i} (\nabla f_{B_i}(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_i)), \beta^{t-j} (\nabla f_{B_j}(\boldsymbol{\theta}_j) - \nabla f(\boldsymbol{\theta}_j)) \rangle. \end{aligned}$$

Let i and j satisfy $0 \leq j < i \leq t$. Proposition A.1 and Assumptions (A2) and (A3) imply that

$$\begin{aligned} &\mathbb{E} [\langle \nabla f_{B_i}(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_i), \nabla f_{B_j}(\boldsymbol{\theta}_j) - \nabla f(\boldsymbol{\theta}_j) \rangle] \\ &= \mathbb{E}_{\boldsymbol{\xi}_0} \mathbb{E}_{\boldsymbol{\xi}_1} \cdots \mathbb{E}_{\boldsymbol{\xi}_t} [\langle \nabla f_{B_i}(\boldsymbol{\theta}_i) - \mathbb{E}_{\boldsymbol{\xi}_i} [\nabla f_{B_i}(\boldsymbol{\theta}_i)], \nabla f_{B_j}(\boldsymbol{\theta}_j) - \mathbb{E}_{\boldsymbol{\xi}_j} [\nabla f_{B_j}(\boldsymbol{\theta}_j)] \rangle] \\ &= \mathbb{E}_{\boldsymbol{\xi}_0} \mathbb{E}_{\boldsymbol{\xi}_1} \cdots \mathbb{E}_{\boldsymbol{\xi}_i} [\langle \nabla f_{B_i}(\boldsymbol{\theta}_i) - \mathbb{E}_{\boldsymbol{\xi}_i} [\nabla f_{B_i}(\boldsymbol{\theta}_i)], \nabla f_{B_j}(\boldsymbol{\theta}_j) - \mathbb{E}_{\boldsymbol{\xi}_j} [\nabla f_{B_j}(\boldsymbol{\theta}_j)] \rangle] \\ &= \mathbb{E}_{\boldsymbol{\xi}_0} \mathbb{E}_{\boldsymbol{\xi}_1} \cdots \mathbb{E}_{\boldsymbol{\xi}_{i-1}} [\langle \mathbb{E}_{\boldsymbol{\xi}_i} [\nabla f_{B_i}(\boldsymbol{\theta}_i)] - \mathbb{E}_{\boldsymbol{\xi}_i} [\nabla f_{B_i}(\boldsymbol{\theta}_i)], \nabla f_{B_j}(\boldsymbol{\theta}_j) - \mathbb{E}_{\boldsymbol{\xi}_j} [\nabla f_{B_j}(\boldsymbol{\theta}_j)] \rangle] \\ &= 0. \end{aligned}$$

A similar argument as in the case of $j < i$ ensures the above equation also holds for $i < j$. Hence, Proposition A.1 guarantees that, for all $t \in \mathbb{N}$,

$$\begin{aligned}
 & \mathbb{E} \left[\left\| \mathbf{m}_t - (1 - \beta) \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) \right\|^2 \right] \\
 &= (1 - \beta)^2 \sum_{i=0}^t \mathbb{E} [\langle \beta^{t-i} (\nabla f_{B_i}(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_i)), \beta^{t-i} (\nabla f_{B_i}(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_i)) \rangle] \\
 &= (1 - \beta)^2 \sum_{i=0}^t \beta^{2(t-i)} \mathbb{E} [\|\nabla f_{B_i}(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_i)\|^2] \\
 &\leq (1 - \beta)^2 \sum_{i=0}^t \beta^{2(t-i)} \frac{\sigma^2}{b_i},
 \end{aligned}$$

which completes the proof. \square

A.2. Proofs of Theorems 3.2 and A.2

Using Lemma A.1, we have the following.

Lemma A.2. *Suppose that Assumption 1 holds and $(\boldsymbol{\theta}_t)$ is the sequence generated by Algorithm 1. We define (\mathbf{z}_t) for all $t \in \{0\} \cup \mathbb{N}$ as*

$$\mathbf{z}_t := \begin{cases} \boldsymbol{\theta}_t & (t = 0) \\ \frac{1}{1-\beta} \boldsymbol{\theta}_t - \frac{\beta}{1-\beta} \boldsymbol{\theta}_{t-1} & (t \geq 1). \end{cases}$$

Then, for all $t \in \{0\} \cup \mathbb{N}$,

$$\begin{aligned}
 \mathbb{E}[f(\mathbf{z}_{t+1})] &\leq \mathbb{E}[f(\mathbf{z}_t)] + \eta \left[L \left\{ \left(\frac{\beta}{1-\beta} \right)^2 + \frac{3}{2} \right\} \eta - 1 \right] \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \\
 &\quad + \frac{L\sigma^2\eta^2}{2} \left\{ \beta^2 \sum_{i=0}^{t-1} \frac{\beta^{2(t-1-i)}}{b_i} + \frac{1}{b_t} \right\} \\
 &\quad + \left(\frac{1}{1-\beta} \right)^2 L\eta^2 (1 - \beta^t)^2 \mathbb{E} \left[\left\| \frac{1-\beta}{1-\beta^t} \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2 \right],
 \end{aligned}$$

where $L := \frac{1}{n} \sum_{i \in [n]} L_i$ is the Lipschitz constant of ∇f and we assume that $\sum_{i=0}^{-1} a_i := 0$ for some $a_i \in \mathbb{R}$.

Proof The descent lemma (Beck, 2017, Lemma 5.7) ensures that, for all $t \in \{0\} \cup \mathbb{N}$,

$$\mathbb{E}_{\boldsymbol{\xi}_t}[f(\mathbf{z}_{t+1})] \leq f(\mathbf{z}_t) + \mathbb{E}_{\boldsymbol{\xi}_t}[\langle \nabla f(\mathbf{z}_t), \mathbf{z}_{t+1} - \mathbf{z}_t \rangle] + \frac{L}{2} \mathbb{E}_{\boldsymbol{\xi}_t}[\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2],$$

which, together with $\mathbf{z}_{t+1} = \mathbf{z}_t - \eta \nabla f_{B_t}(\boldsymbol{\theta}_t)$ (Liu et al., 2020, Lemma 3), implies that

$$\mathbb{E}_{\boldsymbol{\xi}_t}[f(\mathbf{z}_{t+1})] \leq f(\mathbf{z}_t) - \underbrace{\eta \mathbb{E}_{\boldsymbol{\xi}_t}[\langle \nabla f(\mathbf{z}_t), \nabla f_{B_t}(\boldsymbol{\theta}_t) \rangle]}_{X_t} + \frac{L\eta^2}{2} \underbrace{\mathbb{E}_{\boldsymbol{\xi}_t}[\|\nabla f_{B_t}(\boldsymbol{\theta}_t)\|^2]}_{Y_t}. \quad (7)$$

From Proposition A.1, we have that

$$X_t = \langle \nabla f(\mathbf{z}_t), \mathbb{E}_{\xi_t}[\nabla f_{B_t}(\boldsymbol{\theta}_t)] \rangle = \langle \nabla f(\mathbf{z}_t), \nabla f(\boldsymbol{\theta}_t) \rangle,$$

which, together with the Cauchy–Schwarz inequality, Young’s inequality, and L -smoothness of f , implies that, for all $\rho > 0$,

$$\begin{aligned} -\eta X_t &= \langle \nabla f(\mathbf{z}_t), -\eta \nabla f(\boldsymbol{\theta}_t) \rangle \\ &= \langle \nabla f(\mathbf{z}_t) - \nabla f(\boldsymbol{\theta}_t), -\eta \nabla f(\boldsymbol{\theta}_t) \rangle - \eta \|\nabla f(\boldsymbol{\theta}_t)\|^2 \\ &\leq (\sqrt{\eta} \|\nabla f(\mathbf{z}_t) - \nabla f(\boldsymbol{\theta}_t)\|)(\sqrt{\eta} \|\nabla f(\boldsymbol{\theta}_t)\|) - \eta \|\nabla f(\boldsymbol{\theta}_t)\|^2 \\ &\leq \frac{\rho\eta}{2} \|\nabla f(\mathbf{z}_t) - \nabla f(\boldsymbol{\theta}_t)\|^2 + \frac{\eta}{2\rho} \|\nabla f(\boldsymbol{\theta}_t)\|^2 - \eta \|\nabla f(\boldsymbol{\theta}_t)\|^2 \\ &\leq \frac{\rho\eta L^2}{2} \|\mathbf{z}_t - \boldsymbol{\theta}_t\|^2 + \eta \left(\frac{1}{2\rho} - 1 \right) \|\nabla f(\boldsymbol{\theta}_t)\|^2. \end{aligned}$$

The definitions of \mathbf{z}_t and $\boldsymbol{\theta}_t$ ($= \boldsymbol{\theta}_{t-1} - \eta \mathbf{m}_{t-1}$) ensure that, for all $t \geq 1$,

$$\mathbf{z}_t = \frac{1}{1-\beta} \boldsymbol{\theta}_t - \frac{\beta}{1-\beta} (\boldsymbol{\theta}_t + \eta \mathbf{m}_{t-1}) = \boldsymbol{\theta}_t - \frac{\beta}{1-\beta} \eta \mathbf{m}_{t-1}.$$

From $\mathbf{m}_{-1} := \mathbf{0}$ and $\mathbf{z}_0 = \boldsymbol{\theta}_0$, we have that, for all $t \in \{0\} \cup \mathbb{N}$,

$$\mathbf{z}_t = \boldsymbol{\theta}_t - \frac{\beta}{1-\beta} \eta \mathbf{m}_{t-1}.$$

Accordingly, we have that

$$-\eta X_t \leq \frac{\rho\eta^3 L^2}{2} \left(\frac{\beta}{1-\beta} \right)^2 \|\mathbf{m}_{t-1}\|^2 + \eta \left(\frac{1}{2\rho} - 1 \right) \|\nabla f(\boldsymbol{\theta}_t)\|^2. \quad (8)$$

From $\|\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2\|^2 \leq 2\|\boldsymbol{\theta}_1\|^2 + 2\|\boldsymbol{\theta}_2\|^2$, we have that

$$\|\mathbf{m}_{t-1}\|^2 \leq 2 \left\| \mathbf{m}_{t-1} - (1-\beta) \sum_{i=0}^{t-1} \beta^{t-1-i} \nabla f(\boldsymbol{\theta}_i) \right\|^2 + 2 \underbrace{\left\| (1-\beta) \sum_{i=0}^{t-1} \beta^{t-1-i} \nabla f(\boldsymbol{\theta}_i) \right\|^2}_{Z_t}. \quad (9)$$

Moreover, for all $t \geq 2$,

$$\frac{1}{(1-\beta^{t-1})^2} Z_t \leq 2\|\nabla f(\boldsymbol{\theta}_t)\|^2 + 2 \underbrace{\left\| \frac{1-\beta}{1-\beta^{t-1}} \sum_{i=0}^{t-1} \beta^{t-1-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2}_{W_t}. \quad (10)$$

Meanwhile, we also have that

$$\begin{aligned}
 & \left\| \frac{1-\beta}{1-\beta^t} \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2 \\
 &= \left\| \frac{1-\beta}{1-\beta^t} \left(\beta^t \nabla f(\boldsymbol{\theta}_0) + \beta^{t-1} \nabla f(\boldsymbol{\theta}_1) + \dots + \beta^{t-(t-1)} \nabla f(\boldsymbol{\theta}_{t-1}) + \nabla f(\boldsymbol{\theta}_t) \right) - \nabla f(\boldsymbol{\theta}_t) \right\|^2 \\
 &= \left\| \frac{1-\beta}{1-\beta^t} \left(\beta^t \nabla f(\boldsymbol{\theta}_0) + \beta^{t-1} \nabla f(\boldsymbol{\theta}_1) + \dots + \beta^{t-(t-1)} \nabla f(\boldsymbol{\theta}_{t-1}) \right) + \left(\frac{1-\beta}{1-\beta^t} - 1 \right) \nabla f(\boldsymbol{\theta}_t) \right\|^2 \\
 &= \left\| \frac{1-\beta}{1-\beta^t} \left(\beta^t \nabla f(\boldsymbol{\theta}_0) + \beta^{t-1} \nabla f(\boldsymbol{\theta}_1) + \dots + \beta^{t-(t-1)} \nabla f(\boldsymbol{\theta}_{t-1}) \right) - \frac{\beta - \beta^t}{1 - \beta^t} \nabla f(\boldsymbol{\theta}_t) \right\|^2 \\
 &= \left\| \frac{1-\beta}{1-\beta^t} \beta \left(\beta^{t-1} \nabla f(\boldsymbol{\theta}_0) + \dots + \beta^{t-(t-1)} \nabla f(\boldsymbol{\theta}_{t-2}) + \nabla f(\boldsymbol{\theta}_{t-1}) \right) - \frac{1 - \beta^{t-1}}{1 - \beta^t} \beta \nabla f(\boldsymbol{\theta}_t) \right\|^2 \\
 &= \beta^2 \left(\frac{1 - \beta^{t-1}}{1 - \beta^t} \right)^2 \left\| \frac{1 - \beta}{1 - \beta^{t-1}} \sum_{i=0}^{t-1} \beta^{t-1-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2 \\
 &= \beta^2 \left(\frac{1 - \beta^{t-1}}{1 - \beta^t} \right)^2 W_t,
 \end{aligned}$$

which implies that, for all $t \geq 2$,

$$W_t = \frac{(1 - \beta^t)^2}{\beta^2(1 - \beta^{t-1})^2} \left\| \frac{1 - \beta}{1 - \beta^t} \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2. \quad (11)$$

From (9), (10), and (11),

$$\begin{aligned}
 \|\mathbf{m}_{t-1}\|^2 &\leq 2 \left\| \mathbf{m}_{t-1} - (1 - \beta) \sum_{i=0}^{t-1} \beta^{t-1-i} \nabla f(\boldsymbol{\theta}_i) \right\|^2 + 2 \left\{ 2(1 - \beta^{t-1})^2 \|\nabla f(\boldsymbol{\theta}_t)\|^2 \right. \\
 &\quad \left. + 2(1 - \beta^{t-1})^2 \frac{(1 - \beta^t)^2}{\beta^2(1 - \beta^{t-1})^2} \left\| \frac{1 - \beta}{1 - \beta^t} \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2 \right\} \\
 &= 2 \left\| \mathbf{m}_{t-1} - (1 - \beta) \sum_{i=0}^{t-1} \beta^{t-1-i} \nabla f(\boldsymbol{\theta}_i) \right\|^2 + 4(1 - \beta^{t-1})^2 \|\nabla f(\boldsymbol{\theta}_t)\|^2 \\
 &\quad + \frac{4(1 - \beta^t)^2}{\beta^2} \left\| \frac{1 - \beta}{1 - \beta^t} \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2
 \end{aligned}$$

Hence, (8) ensures that

$$\begin{aligned}
-\eta X_t &\leq \frac{\rho\eta^3 L^2}{2} \left(\frac{\beta}{1-\beta} \right)^2 \left\{ 2 \left\| \mathbf{m}_{t-1} - (1-\beta) \sum_{i=0}^{t-1} \beta^{t-1-i} \nabla f(\boldsymbol{\theta}_i) \right\|^2 + 4(1-\beta^{t-1})^2 \|\nabla f(\boldsymbol{\theta}_t)\|^2 \right. \\
&\quad \left. + \frac{4(1-\beta^t)^2}{\beta^2} \left\| \frac{1-\beta}{1-\beta^t} \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2 \right\} + \eta \left(\frac{1}{2\rho} - 1 \right) \|\nabla f(\boldsymbol{\theta}_t)\|^2 \\
&= \rho\eta^3 L^2 \left(\frac{\beta}{1-\beta} \right)^2 \left\| \mathbf{m}_{t-1} - (1-\beta) \sum_{i=0}^{t-1} \beta^{t-1-i} \nabla f(\boldsymbol{\theta}_i) \right\|^2 \\
&\quad + 2\rho\eta^3 L^2 \left(\frac{\beta}{1-\beta} \right)^2 (1-\beta^{t-1})^2 \|\nabla f(\boldsymbol{\theta}_t)\|^2 \\
&\quad + 2\rho\eta^3 L^2 \left(\frac{1}{1-\beta} \right)^2 (1-\beta^t)^2 \left\| \frac{1-\beta}{1-\beta^t} \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2 \\
&\quad + \eta \left(\frac{1}{2\rho} - 1 \right) \|\nabla f(\boldsymbol{\theta}_t)\|^2.
\end{aligned}$$

Let us take the total expectation on both sides of the above inequality. Lemma A.1 then guarantees that, for all $t \geq 2$ and for all $\rho > 0$,

$$\begin{aligned}
-\eta \mathbb{E}[X_t] &\leq \rho\eta^3 L^2 \left(\frac{\beta}{1-\beta} \right)^2 (1-\beta)^2 \sigma^2 \sum_{i=0}^{t-1} \frac{\beta^{2(t-1-i)}}{b_i} \\
&\quad + 2\rho\eta^3 L^2 \left(\frac{\beta}{1-\beta} \right)^2 (1-\beta^{t-1})^2 \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \\
&\quad + 2\rho\eta^3 L^2 \left(\frac{1}{1-\beta} \right)^2 (1-\beta^t)^2 \mathbb{E} \left[\left\| \frac{1-\beta}{1-\beta^t} \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2 \right] \\
&\quad + \eta \left(\frac{1}{2\rho} - 1 \right) \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \tag{12} \\
&\leq \rho\eta^3 L^2 \beta^2 \sigma^2 \sum_{i=0}^{t-1} \frac{\beta^{2(t-1-i)}}{b_i} + 2\rho\eta^3 L^2 \left(\frac{\beta}{1-\beta} \right)^2 \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \\
&\quad + 2\rho\eta^3 L^2 \left(\frac{1}{1-\beta} \right)^2 (1-\beta^t)^2 \mathbb{E} \left[\left\| \frac{1-\beta}{1-\beta^t} \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2 \right] \\
&\quad + \eta \left(\frac{1}{2\rho} - 1 \right) \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2].
\end{aligned}$$

Moreover, Proposition A.1 guarantees that

$$\begin{aligned}
 \mathbb{E}[Y_t] &= \mathbb{E}_{\boldsymbol{\xi}_t} \left[\|\nabla f_{B_t}(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t) + \nabla f(\boldsymbol{\theta}_t)\|^2 \middle| \hat{\boldsymbol{\xi}}_{t-1} \right] \\
 &= \mathbb{E}[\|\nabla f_{B_t}(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t)\|^2] + 2\mathbb{E}[\langle \nabla f_{B_t}(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t), \nabla f(\boldsymbol{\theta}_t) \rangle] + \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \\
 &\leq \frac{\sigma^2}{b_t} + \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2].
 \end{aligned} \tag{13}$$

Therefore, from (7), (12), and (13), for all $t \geq 2$ and for all $\rho > 0$, we have

$$\begin{aligned}
 \mathbb{E}[f(\mathbf{z}_{t+1})] &\leq \mathbb{E}[f(\mathbf{z}_t)] - \eta \mathbb{E}[X_t] + \frac{L\eta^2}{2} \mathbb{E}[Y_t] \\
 &\leq \mathbb{E}[f(\mathbf{z}_t)] + \rho\eta^3 L^2 \beta^2 \sigma^2 \sum_{i=0}^{t-1} \frac{\beta^{2(t-1-i)}}{b_i} \\
 &\quad + 2\rho\eta^3 L^2 \left(\frac{\beta}{1-\beta} \right)^2 \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] + \eta \left(\frac{1}{2\rho} - 1 \right) \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \\
 &\quad + 2\rho\eta^3 L^2 \left(\frac{1}{1-\beta} \right)^2 (1-\beta^t)^2 \left\| \frac{1-\beta}{1-\beta^t} \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2 \\
 &\quad + \frac{L\eta^2}{2} \left(\frac{\sigma^2}{b_t} + \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \right) \\
 &= \mathbb{E}[f(\mathbf{z}_t)] + \underbrace{\left\{ 2\rho\eta^3 L^2 \left(\frac{\beta}{1-\beta} \right)^2 + \eta \left(\frac{1}{2\rho} - 1 \right) + \frac{L\eta^2}{2} \right\}}_A \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \\
 &\quad + \underbrace{L\eta^2 \sigma^2 \left(\rho\eta L \beta^2 \sum_{i=0}^{t-1} \frac{\beta^{2(t-1-i)}}{b_i} + \frac{1}{2b_t} \right)}_{B_t} \\
 &\quad + \underbrace{2\rho\eta^3 L^2 \left(\frac{1}{1-\beta} \right)^2 (1-\beta^t)^2 \left\| \frac{1-\beta}{1-\beta^t} \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2}_C.
 \end{aligned}$$

The setting $\rho := \frac{1}{2L\eta}$ implies that, for all $t \geq 2$,

$$\begin{aligned}
 A &= \frac{L^2\eta^3}{L\eta} \left(\frac{\beta}{1-\beta} \right)^2 + \eta(L\eta - 1) + \frac{L\eta^2}{2} = L\eta^2 \left(\frac{\beta}{1-\beta} \right)^2 + \eta(L\eta - 1) + \frac{L\eta^2}{2} \\
 &= L \left\{ \left(\frac{\beta}{1-\beta} \right)^2 + \frac{3}{2} \right\} \eta^2 - \eta, \\
 B_t &= \frac{L\eta\beta^2}{2L\eta} \sum_{i=0}^{t-1} \frac{\beta^{2(t-1-i)}}{b_i} + \frac{1}{2b_t} = \frac{1}{2} \left(\beta^2 \sum_{i=0}^{t-1} \frac{\beta^{2(t-1-i)}}{b_i} + \frac{1}{b_t} \right), \quad C = \frac{2L^2\eta^3}{2L\eta} = L\eta^2.
 \end{aligned}$$

When $t = 0$, from $\|\mathbf{m}_{-1}\| = 0$ and $\sum_{i=0}^{-1} a_i := 0$, the assertion in Lemma A.2 holds. When $t = 1$, assuming $\frac{1}{1-\beta^0} := 1$, the assertion in Lemma A.2 again holds. This completes the proof. \square

Using Lemma A.2, we have the following lemma.

Lemma A.3. *Suppose that Assumption 1 holds and $(\boldsymbol{\theta}_t)$ is the sequence generated by Algorithm 1 with $\eta > 0$ satisfying*

$$\eta \leq \frac{1 - \beta}{2\sqrt{2}\sqrt{\beta + \beta^2}L}.$$

Let (\mathbf{z}_t) be the sequence defined as in Lemma A.2 and define $L_t \in \mathbb{R}$ for all $t \in \{0\} \cup \mathbb{N}$ as

$$L_t := \begin{cases} f(\mathbf{z}_0) - f^* & (t = 0) \\ f(\mathbf{z}_1) - f^* & (t = 1) \\ f(\mathbf{z}_t) - f^* + \sum_{i=1}^{t-1} c_i \|\boldsymbol{\theta}_{t+1-i} - \boldsymbol{\theta}_{t-i}\|^2 & (t \geq 2), \end{cases}$$

where f^ is the minimum value of f over \mathbb{R}^d and $(c_i) \subset \mathbb{R}_{++}$ is defined by*

$$c_1 = \frac{(\beta + \beta^2)L^3\eta^2}{(1 - \beta)^2\{(1 - \beta)^2 - 4(\beta + \beta^2)L^2\eta^2\}} \text{ and } c_{i+1} = c_i - \left(4c_1\eta^2 + \frac{L\eta^2}{(1 - \beta)^2}\right)\beta^i \left(i + \frac{\beta}{1 - \beta}\right)L^2 \quad (i \in [t - 1]).$$

Then, for all $t \in \{0\} \cup \mathbb{N}$,

$$\begin{aligned} \mathbb{E}[L_{t+1} - L_t] &\leq \eta \left[L \left\{ \left(\frac{\beta}{1 - \beta} \right)^2 + \frac{3}{2} \right\} \eta - 1 + 4c_1\eta \right] \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \\ &\quad + \frac{L\sigma^2\eta^2}{2} \left\{ \beta^2 \sum_{i=0}^{t-1} \frac{\beta^{2(t-1-i)}}{b_i} + \frac{1}{b_t} \right\} + 2c_1\eta^2(1 - \beta)^2\sigma^2 \sum_{i=0}^t \frac{\beta^{2(t-i)}}{b_i}, \end{aligned}$$

where we assume that $\sum_{i=0}^{-1} a_i := 0$ for some $a_i \in \mathbb{R}$.

Proof Let $t \geq 2$. The definition of L_t implies that

$$\begin{aligned} \mathbb{E}[L_{t+1} - L_t] &= \mathbb{E}[f(\mathbf{z}_{t+1}) - f(\mathbf{z}_t)] + \mathbb{E} \left[\sum_{i=1}^t c_i \|\boldsymbol{\theta}_{t+2-i} - \boldsymbol{\theta}_{t+1-i}\|^2 - \sum_{i=1}^{t-1} c_i \|\boldsymbol{\theta}_{t+1-i} - \boldsymbol{\theta}_{t-i}\|^2 \right] \\ &= \mathbb{E}[f(\mathbf{z}_{t+1}) - f(\mathbf{z}_t)] + \mathbb{E}[c_1 \|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|^2] \\ &\quad + \mathbb{E} \left[\sum_{i=1}^{t-1} c_{i+1} \|\boldsymbol{\theta}_{t+1-i} - \boldsymbol{\theta}_{t-i}\|^2 - \sum_{i=1}^{t-1} c_i \|\boldsymbol{\theta}_{t+1-i} - \boldsymbol{\theta}_{t-i}\|^2 \right] \\ &= \mathbb{E}[f(\mathbf{z}_{t+1}) - f(\mathbf{z}_t)] + \mathbb{E}[c_1 \|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|^2] \\ &\quad + \sum_{i=1}^{t-1} (c_{i+1} - c_i) \mathbb{E}[\|\boldsymbol{\theta}_{t+1-i} - \boldsymbol{\theta}_{t-i}\|^2]. \end{aligned}$$

Lemma A.2 thus ensures that

$$\begin{aligned} \mathbb{E}[L_{t+1} - L_t] &\leq \eta \left[L \left\{ \left(\frac{\beta}{1 - \beta} \right)^2 + \frac{3}{2} \right\} \eta - 1 \right] \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] + \frac{L\sigma^2\eta^2}{2} \left\{ \beta^2 \sum_{i=0}^{t-1} \frac{\beta^{2(t-1-i)}}{b_i} + \frac{1}{b_t} \right\} \\ &\quad + \left(\frac{1}{1 - \beta} \right)^2 L\eta^2(1 - \beta^t)^2 \mathbb{E} \left[\left\| \frac{1 - \beta}{1 - \beta^t} \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2 \right] \\ &\quad + c_1\eta^2 \mathbb{E}[\|\mathbf{m}_t\|^2] + \sum_{i=1}^{t-1} (c_{i+1} - c_i) \mathbb{E}[\|\boldsymbol{\theta}_{t+1-i} - \boldsymbol{\theta}_{t-i}\|^2]. \end{aligned} \tag{14}$$

A similar discussion to the one showing (9) and (10) ensures that

$$\begin{aligned} \|\mathbf{m}_t\|^2 &\leq 2 \left\| \mathbf{m}_t - (1 - \beta) \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) \right\|^2 \\ &\quad + 2 \left\{ 2(1 - \beta^t)^2 \|\nabla f(\boldsymbol{\theta}_t)\|^2 + 2(1 - \beta^t)^2 \left\| \frac{1 - \beta}{1 - \beta^t} \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2 \right\}, \end{aligned}$$

which, together with Lemma A.1 and $1 - \beta^t \leq 1$, implies that

$$\begin{aligned} \mathbb{E}[\|\mathbf{m}_t\|^2] &\leq 2(1 - \beta)^2 \sigma^2 \sum_{i=0}^t \frac{\beta^{2(t-i)}}{b_i} + 4(1 - \beta^t)^2 \|\nabla f(\boldsymbol{\theta}_t)\|^2 \\ &\quad + 4(1 - \beta^t)^2 \left\| \frac{1 - \beta}{1 - \beta^t} \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2 \\ &\leq 2(1 - \beta)^2 \sigma^2 \sum_{i=0}^t \frac{\beta^{2(t-i)}}{b_i} + 4\|\nabla f(\boldsymbol{\theta}_t)\|^2 \\ &\quad + 4(1 - \beta^t)^2 \underbrace{\left\| \frac{1 - \beta}{1 - \beta^t} \sum_{i=0}^t \beta^{t-i} \nabla f(\boldsymbol{\theta}_i) - \nabla f(\boldsymbol{\theta}_t) \right\|^2}_{V_t}. \end{aligned}$$

Lemma 2 in (Liu et al., 2020) guarantees that

$$\mathbb{E}[V_t] \leq \sum_{i=1}^{t-1} a_{t,i} \mathbb{E}[\|\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i\|^2], \text{ where } a_{t,i} := \frac{L^2 \beta^{t-i}}{1 - \beta^t} \left(t - i + \frac{\beta}{1 - \beta} \right),$$

which implies that

$$\mathbb{E}[V_t] \leq \sum_{i=1}^{t-1} a_{t,t-i} \mathbb{E}[\|\boldsymbol{\theta}_{t+1-i} - \boldsymbol{\theta}_{t-i}\|^2], \text{ where } a_{t,t-i} := \frac{L^2 \beta^i}{1 - \beta^t} \left(i + \frac{\beta}{1 - \beta} \right). \quad (15)$$

Moreover, $c_1 > 0$ when $\eta \leq \frac{1-\beta}{2\sqrt{2}L\sqrt{\beta+\beta^2}}$. Hence, (14) ensures that

$$\begin{aligned}
& \mathbb{E}[L_{t+1} - L_t] \\
& \leq \eta \left[L \left\{ \left(\frac{\beta}{1-\beta} \right)^2 + \frac{3}{2} \right\} \eta - 1 \right] \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \\
& \quad + \frac{L\sigma^2\eta^2}{2} \left\{ \beta^2 \sum_{i=0}^{t-1} \frac{\beta^{2(t-1-i)}}{b_i} + \frac{1}{b_t} \right\} + \left(\frac{1}{1-\beta} \right)^2 L\eta^2(1-\beta^t)^2 \mathbb{E}[V_t] \\
& \quad + c_1\eta^2 \left\{ 2(1-\beta)^2\sigma^2 \sum_{i=0}^t \frac{\beta^{2(t-i)}}{b_i} + 4\mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] + 4(1-\beta^t)^2 \mathbb{E}[V_t] \right\} \\
& \quad + \sum_{i=1}^{t-1} (c_{i+1} - c_i) \mathbb{E}[\|\boldsymbol{\theta}_{t+1-i} - \boldsymbol{\theta}_{t-i}\|^2] \\
& = \eta \left[L \left\{ \left(\frac{\beta}{1-\beta} \right)^2 + \frac{3}{2} \right\} \eta - 1 + 4c_1\eta \right] \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \\
& \quad + \frac{L\sigma^2\eta^2}{2} \left\{ \beta^2 \sum_{i=0}^{t-1} \frac{\beta^{2(t-1-i)}}{b_i} + \frac{1}{b_t} \right\} + 2c_1\eta^2(1-\beta)^2\sigma^2 \sum_{i=0}^t \frac{\beta^{2(t-i)}}{b_i} \\
& \quad + \underbrace{\sum_{i=1}^{t-1} \left\{ \left(\frac{1}{1-\beta} \right)^2 L\eta^2(1-\beta^t)^2 a_{t,t-i} + 4c_1\eta^2(1-\beta^t)^2 a_{t,t-i} + (c_{i+1} - c_i) \right\}}_{N_{t,i}} \mathbb{E}[\|\boldsymbol{\theta}_{t+1-i} - \boldsymbol{\theta}_{t-i}\|^2].
\end{aligned}$$

Finally, we prove that $N_{t,i} \leq 0$. From the definition of $a_{t,t-i}$ in (15) and

$$c_{i+1} = c_i - \left(4c_1\eta^2 + \frac{L\eta^2}{(1-\beta)^2} \right) \beta^i \left(i + \frac{\beta}{1-\beta} \right) L^2,$$

we have that

$$\begin{aligned}
N_{t,i} &= \left(\frac{1}{1-\beta} \right)^2 L\eta^2(1-\beta^t)^2 a_{t,t-i} + 4c_1\eta^2(1-\beta^t)^2 a_{t,t-i} + (c_{i+1} - c_i) \\
&= \left\{ \left(\frac{1}{1-\beta} \right)^2 L\eta^2(1-\beta^t)^2 + 4c_1\eta^2(1-\beta^t)^2 \right\} \frac{L^2\beta^i}{1-\beta^t} \left(i + \frac{\beta}{1-\beta} \right) \\
&\quad - \left(4c_1\eta^2 + \frac{L\eta^2}{(1-\beta)^2} \right) \beta^i \left(i + \frac{\beta}{1-\beta} \right) L^2 \\
&= L^2\eta^2\beta^i \left(i + \frac{\beta}{1-\beta} \right) \left[\left\{ \frac{1-\beta^t}{(1-\beta)^2} L + 4c_1(1-\beta^t) \right\} - \left\{ 4c_1 + \frac{L}{(1-\beta)^2} \right\} \right] \\
&= L^2\eta^2\beta^i \left(i + \frac{\beta}{1-\beta} \right) \left[\frac{L}{(1-\beta)^2} (1-\beta^t - 1) + 4c_1(1-\beta^t - 1) \right] \\
&= -L^2\eta^2\beta^i \left(i + \frac{\beta}{1-\beta} \right) \left[\frac{L}{(1-\beta)^2} + 4c_1 \right] \beta^t.
\end{aligned}$$

From

$$\eta \leq \frac{1-\beta}{2\sqrt{2}L\sqrt{\beta+\beta^2}} \text{ and } c_1 = \frac{(\beta+\beta^2)L^3\eta^2}{(1-\beta)^2\{(1-\beta)^2-4(\beta+\beta^2)L^2\eta^2\}}, \quad (16)$$

we have that

$$c_1 = \frac{\eta^2 L^3 \frac{\beta+\beta^2}{(1-\beta)^4}}{1-4\eta^2 L^2 \frac{\beta+\beta^2}{(1-\beta)^2}} > 0.$$

Accordingly,

$$N_{t,i} = -L^2\eta^2\beta^i \left(i + \frac{\beta}{1-\beta} \right) \left[\frac{L}{(1-\beta)^2} + 4c_1 \right] \beta^t < 0.$$

This completes the proof. \square

Lemma A.3 leads to the following.

Lemma A.4. *Suppose that Assumption 1 holds and $(\boldsymbol{\theta}_t)$ is the sequence generated by Algorithm 1 with $\eta > 0$ satisfying*

$$\eta \leq \max \left\{ \frac{1-\beta}{2\sqrt{2}\sqrt{\beta+\beta^2}L}, \frac{(1-\beta)^2}{(5\beta^2-6\beta+5)L} \right\}.$$

Then, for all $T \geq 1$,

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \leq \frac{2(f(\boldsymbol{\theta}_0) - f^*)}{\eta T} + \frac{2L\eta\sigma^2}{T} \sum_{t=0}^{T-1} \sum_{i=0}^t \frac{\beta^{2(t-i)}}{b_i}.$$

Proof Lemma A.3 guarantees that, for all $t \in \{0\} \cup \mathbb{N}$,

$$\begin{aligned} & \underbrace{-\eta \left[L \left\{ \left(\frac{\beta}{1-\beta} \right)^2 + \frac{3}{2} \right\} \eta - 1 + 4c_1\eta \right] \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2]}_D \\ & \leq \mathbb{E}[L_{t+1} - L_t] + \underbrace{\frac{L\sigma^2\eta^2}{2} \left\{ \beta^2 \sum_{i=0}^{t-1} \frac{\beta^{2(t-1-i)}}{b_i} + \frac{1}{b_t} \right\} + 2c_1\eta^2(1-\beta)^2\sigma^2 \sum_{i=0}^t \frac{\beta^{2(t-i)}}{b_i}}_{U_t}, \end{aligned}$$

where $\beta \in [0, 1)$, and η and c_1 satisfy (16). Summing the above inequality from $t = 0$ to $t = T - 1$ gives

$$D \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \leq \sum_{t=0}^{T-1} \mathbb{E}[L_t - L_{t+1}] + \sum_{t=0}^{T-1} U_t = \mathbb{E}[L_0 - L_T] + \sum_{t=0}^{T-1} U_t,$$

which, together with $L_T \geq 0$, implies that

$$D \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \leq L_0 + \sum_{t=0}^{T-1} U_t. \quad (17)$$

From (16), we have

$$c_1 = \frac{\eta^2 L^3 \frac{\beta + \beta^2}{(1-\beta)^4}}{1 - 4\eta^2 L^2 \frac{\beta + \beta^2}{(1-\beta)^2}} \text{ and } \eta^2 L^2 \frac{\beta + \beta^2}{(1-\beta)^2} \leq \frac{1}{8},$$

which implies that

$$c_1 \leq \frac{L}{8(1-\beta)^2} \left(1 - \frac{4}{8}\right)^{-1} = \frac{L}{4(1-\beta)^2}. \quad (18)$$

Accordingly, from (18) and $\eta \leq \frac{(1-\beta)^2}{L(5\beta^2 - 6\beta + 5)}$, we have that

$$\begin{aligned} D &= -L \left\{ \left(\frac{\beta}{1-\beta} \right)^2 + \frac{3}{2} \right\} \eta^2 + \eta - 4c_1 \eta^2 \geq -L \left\{ \left(\frac{\beta}{1-\beta} \right)^2 + \frac{3}{2} \right\} \eta^2 + \eta - \frac{L\eta^2}{(1-\beta)^2} \\ &= -L\eta^2 \frac{5\beta^2 - 6\beta + 5}{2(1-\beta)^2} + \eta \geq -\frac{\eta}{2} + \eta = \frac{\eta}{2} > 0. \end{aligned} \quad (19)$$

Moreover, from (18), we have

$$\begin{aligned} U_t &= \frac{L\sigma^2\eta^2}{2} \left\{ \beta^2 \sum_{i=0}^{t-1} \frac{\beta^{2(t-1-i)}}{b_i} + \frac{1}{b_t} \right\} + 2c_1\eta^2(1-\beta)^2\sigma^2 \sum_{i=0}^t \frac{\beta^{2(t-i)}}{b_i} \\ &= \sigma^2 \left\{ \frac{L\eta^2}{2} + 2c_1\eta^2(1-\beta)^2 \right\} \sum_{i=0}^t \frac{\beta^{2(t-i)}}{b_i} \\ &\leq \sigma^2 \left(\frac{L\eta^2}{2} + \frac{L\eta^2}{2} \right) \sum_{i=0}^t \frac{\beta^{2(t-i)}}{b_i} = L\eta^2\sigma^2 \sum_{i=0}^t \frac{\beta^{2(t-i)}}{b_i}. \end{aligned} \quad (20)$$

Therefore, (17), (19), and (20) ensure that

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \leq \frac{L_0}{DT} + \frac{1}{T} \sum_{t=0}^{T-1} U_t \leq \frac{2L_0}{\eta T} + \frac{2L\eta\sigma^2}{T} \sum_{t=0}^{T-1} \sum_{i=0}^t \frac{\beta^{2(t-i)}}{b_i}.$$

This completes the proof. \square

Proof [Theorem 3.2]: Let b_t be defined by (4), i.e., for all $m \in [0 : M]$ and all $t \in S_m = \mathbb{N} \cap [\sum_{k=0}^{m-1} K_k E_k, \sum_{k=0}^m K_k E_k)$ ($S_0 := \mathbb{N} \cap [0, K_0 E_0)$),

$$b_t = \delta^m \left\lceil \frac{t}{\sum_{k=0}^m K_k E_k} \right\rceil b_0,$$

which implies that $b_j = \delta^j b_0$ ($j \in S_j$) and

$$(b_0, b_1, \dots, b_M) = (\underbrace{b_0, b_0, \dots, b_0}_{K_0 E_0}, \underbrace{\delta b_0, \delta b_0, \dots, \delta b_0}_{K_1 E_1}, \dots, \underbrace{\delta^M b_0, \delta^M b_0, \dots, \delta^M b_0}_{K_M E_M}),$$

where $T = \sum_{m=0}^M K_m E_m$. Define $K_{\max} := \max\{K_m : m \in [0 : M]\}$ and $E_{\max} := \max\{E_m : m \in [0 : M]\}$. Then, we have

$$\begin{aligned} \sum_{t=0}^{T-1} \sum_{i=0}^t \frac{\beta^{2(t-i)}}{b_i} &= \sum_{t=0}^{T-1} \sum_{i=0}^t \beta^{2(t-i)} \frac{K_i E_i}{\delta^i b_0} \leq \frac{K_{\max} E_{\max}}{b_0} \sum_{t=0}^{T-1} \sum_{i=0}^t \frac{\beta^{2(t-i)}}{\delta^i} \\ &= \frac{K_{\max} E_{\max}}{b_0} \sum_{t=0}^{T-1} \beta^{2t} \sum_{i=0}^t \frac{1}{(\beta^2 \delta)^i} = \frac{K_{\max} E_{\max}}{b_0} \sum_{t=0}^{T-1} \beta^{2t} \frac{1 - (\frac{1}{\beta^2 \delta})^{t+1}}{1 - \frac{1}{\beta^2 \delta}} \\ &= \frac{K_{\max} E_{\max}}{b_0} \sum_{t=0}^{T-1} \beta^{2t} \frac{1 - (\frac{1}{\beta^2 \delta})^{t+1}}{1 - \frac{1}{\beta^2 \delta}} = \frac{K_{\max} E_{\max} \delta}{b_0 (\beta^2 \delta - 1)} \sum_{t=0}^{T-1} \left\{ \beta^{2(t+1)} - \frac{1}{\delta^{t+1}} \right\}, \end{aligned}$$

which implies that

$$\sum_{t=0}^{T-1} \sum_{i=0}^t \frac{\beta^{2(t-i)}}{b_i} \leq \frac{K_{\max} E_{\max} \delta}{b_0 (\beta^2 \delta - 1)} \left\{ \frac{\beta^2 (1 - \beta^{2T})}{1 - \beta^2} - \frac{1 - (\frac{1}{\delta})^T}{\delta - 1} \right\} \leq \frac{K_{\max} E_{\max} \delta}{b_0 (\beta^2 \delta - 1)} \left(\frac{\beta^2}{1 - \beta^2} - \frac{1}{\delta - 1} \right).$$

This completes the proof. \square

Proof [Theorem A.2]: NSHB with $\eta = \frac{\alpha}{1-\beta}$ coincides with SHB defined by (2) (Section 2.2). Hence, Theorem 3.2 leads to Theorem A.2. \square

A.3. Proofs of Theorems 3.1 and A.1

Lemma A.1 and the proof of Lemma A.2 with $\rho = \frac{1-\beta}{2L\eta}$ immediately yield the following lemma. Hence, we will omit its proof.

Lemma A.5. *Suppose that Assumption 1 holds and (θ_t) is the sequence generated by Algorithm 1. Let (z_t) be the sequence defined as in Lemma A.2. Then, for all $t \in \{0\} \cup \mathbb{N}$,*

$$\begin{aligned} \mathbb{E}[f(z_{t+1})] &\leq \mathbb{E}[f(z_t)] + \eta \left\{ L \left(\frac{1 + \beta^2}{1 - \beta} + \frac{1}{2} \right) \eta - 1 \right\} \mathbb{E}[\|\nabla f(\theta_t)\|^2] \\ &\quad + \frac{L\sigma^2\eta^2}{2} \left(\frac{\beta^2}{1 + \beta} + 1 \right) + \frac{(1 - \beta^t)^2}{1 - \beta} L\eta^2 \mathbb{E} \left[\left\| \frac{1 - \beta}{1 - \beta^t} \sum_{i=0}^t \beta^{t-i} \nabla f(\theta_i) - \nabla f(\theta_t) \right\|^2 \right], \end{aligned}$$

where $L := \frac{1}{n} \sum_{i \in [n]} L_i$ is the Lipschitz constant of ∇f and we assume that $\sum_{i=0}^{-1} a_i := 0$ for some $a_i \in \mathbb{R}$.

Moreover, Lemma A.5 and the proof of Lemma A.3 with $\rho = \frac{1-\beta}{2L\eta}$ yield the following lemma, whose proof we will also omit.

Lemma A.6. *Suppose that Assumption 1 holds and (θ_t) is the sequence generated by Algorithm 1 with $\eta > 0$ satisfying*

$$\eta \leq \frac{1 - \beta}{2\sqrt{2}\sqrt{\beta + \beta^2}L}.$$

Let (z_t) be the sequence defined as in Lemma A.2 and let $L_t \in \mathbb{R}$ be defined as in Lemma A.3, where $(c_i) \subset \mathbb{R}_{++}$ is defined by

$c_1 = \frac{(\beta + \beta^2)L^3\eta^2}{(1 - \beta)\{(1 - \beta)^2 - 4(\beta + \beta^2)L^2\eta^2\}}$ and $c_{i+1} = c_i - \left(4c_1\eta^2 + \frac{L\eta^2}{1 - \beta}\right)\beta^i\left(i + \frac{\beta}{1 - \beta}\right)L^2$ ($i \in [t - 1]$).
Then, for all $t \in \{0\} \cup \mathbb{N}$,

$$\begin{aligned} \mathbb{E}[L_{t+1} - L_t] &\leq \eta \left\{ \frac{(3 - \beta + \beta^2)L\eta}{2(1 - \beta)} - 1 + 4c_1\eta \right\} \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \\ &\quad + \left\{ \left(\frac{\beta^2}{1 + \beta} + 1 \right) \frac{L}{2} + \frac{2c_1(1 - \beta)}{1 + \beta} \right\} \frac{\eta^2\sigma^2}{b}. \end{aligned}$$

Lemma A.6 and the proof of Lemma A.4 with $\rho = \frac{1 - \beta}{2L\eta}$ lead to the following.

Lemma A.7. *Suppose that Assumption 1 holds and $(\boldsymbol{\theta}_t)$ is the sequence generated by Algorithm 1 with $\eta > 0$ satisfying*

$$\eta \leq \max \left\{ \frac{1 - \beta}{2\sqrt{2}\sqrt{\beta + \beta^2}L}, \frac{1 - \beta}{(5 - \beta + 2\beta^2)L} \right\}.$$

Then, for all $T \geq 1$,

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \leq \frac{2(f(\boldsymbol{\theta}_0) - f^*)}{\eta T} + \frac{L\eta\sigma^2}{b} \left\{ \frac{\beta + 3\beta^2}{2(1 + \beta)} + 1 \right\}.$$

Proof [Theorems 3.1 and A.1]: Lemma A.7 leads to the assertion in Theorem 3.1. NSHB with $\eta = \frac{\alpha}{1 - \beta}$ coincides with SHB defined by (2) (Section 2.2). Hence, Theorem 3.1 leads to Theorem A.1. \square

Theorem A.1 (Upper bound of $\min_t \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|]$ of mini-batch SHB with Constant BS). *Suppose that Assumption 1 holds and consider the sequence $(\boldsymbol{\theta}_t)$ generated by (2) with a momentum weight $\beta \in (0, 1)$, a constant learning rate $\alpha > 0$ such that*

$$\alpha \leq \frac{(1 - \beta)^2}{2\sqrt{2}\sqrt{\beta + \beta^2}L},$$

and Constant BS defined by (3). Then, for all $T \geq 1$,

$$\min_{t \in [0:T-1]} \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|^2] \leq \frac{2(1 - \beta)(f(\boldsymbol{\theta}_0) - f^*)}{\alpha T} + \frac{L\alpha\sigma^2}{(1 - \beta)b} \left\{ \frac{3\beta^2 + \beta}{2(1 + \beta)} + 1 \right\},$$

that is,

$$\min_{t \in [0:T-1]} \mathbb{E}[\|\nabla f(\boldsymbol{\theta}_t)\|] = O \left(\sqrt{\frac{1}{T} + \frac{\sigma^2}{b}} \right).$$

Theorem A.2 (Convergence of mini-batch SHB with Exponential Growth BS). *Suppose that Assumption 1 holds and consider the sequence $(\boldsymbol{\theta}_t)$ generated by (2) with a momentum weight $\beta \in (0, 1)$, a constant learning rate $\alpha > 0$ such that*

$$\alpha \leq \max \left\{ \frac{(1 - \beta)^2}{2\sqrt{2}\sqrt{\beta + \beta^2}L}, \frac{(1 - \beta)^3}{(5\beta^2 - 6\beta + 5)L} \right\},$$

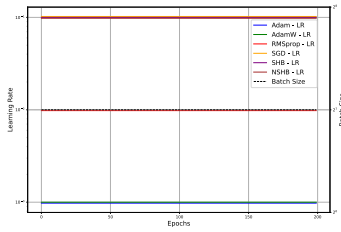
and Exponential Growth BS defined by (4) with $\delta > 1$ and $\beta^2\delta > 1$. Then, for all $T \geq 1$,

$$\min_{t \in [0:T-1]} \mathbb{E}[\|\nabla f(\theta_t)\|^2] \leq \frac{2(1-\beta)(f(\theta_0) - f^*)}{\alpha T} + \frac{2L\alpha\sigma^2 K_{\max} E_{\max} \delta}{(1-\beta)(\beta^2\delta - 1)b_0 T} \left(\frac{\beta^2}{1-\beta^2} - \frac{1}{\delta-1} \right),$$

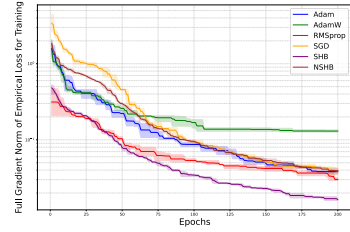
that is,

$$\min_{t \in [0:T-1]} \mathbb{E}[\|\nabla f(\theta_t)\|] = O\left(\frac{1}{\sqrt{T}}\right).$$

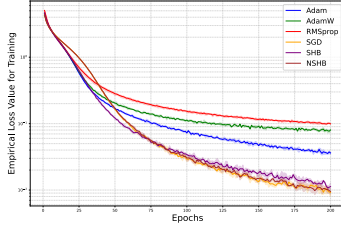
A.4. Training ResNet-18 on Tiny ImageNet



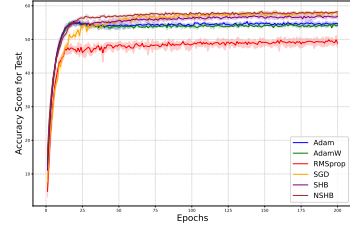
Learning rate and batch size schedules
(a)



Full gradient norm versus epochs
(b)

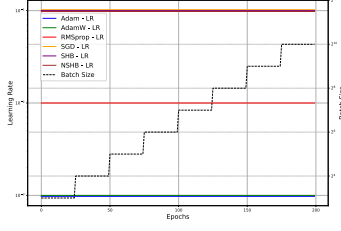


Empirical loss versus epochs
(c)

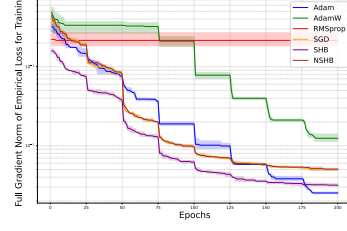


Test accuracy score versus epochs
(d)

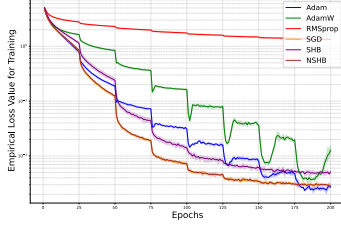
Figure 4: (a) Schedules for each optimizer with constant learning rates and constant batch size, (b) Full gradient norm of empirical loss for training, (c) Empirical loss value for training, and (d) Accuracy score for test to train ResNet-18 on the Tiny ImageNet dataset.



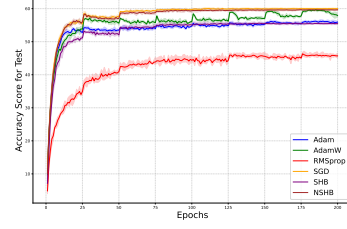
Learning rate and batch size schedules
(a)



Full gradient norm versus epochs
(b)

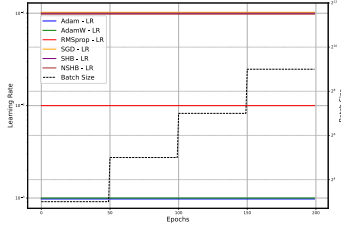


Empirical loss versus epochs
(c)

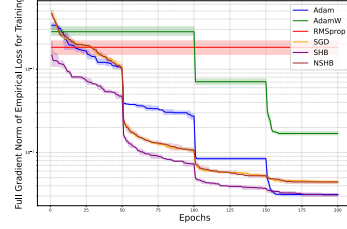


Test accuracy score versus epochs
(d)

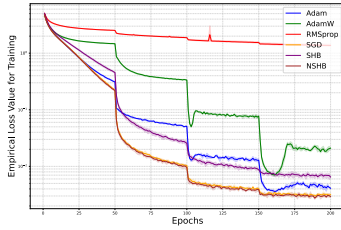
Figure 5: (a) Schedules for each optimizer with constant learning rates and batch size doubling every 25 epochs, (b) Full gradient norm of empirical loss for training, (c) Empirical loss value for training, and (d) Accuracy score for test to train ResNet-18 on the Tiny ImageNet dataset.



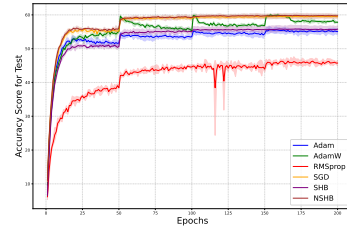
Learning rate and batch size schedules
(a)



Full gradient norm versus epochs
(b)



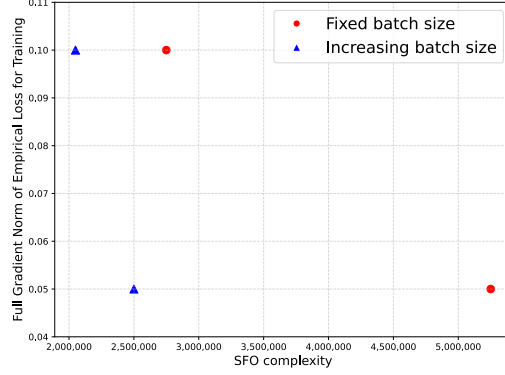
Empirical loss versus epochs
(c)



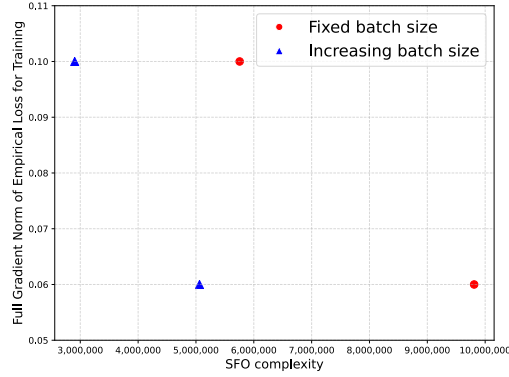
Test accuracy score versus epochs
(d)

Figure 6: (a) Schedules for each optimizer with constant learning rates and batch size quadrupling every 50 epochs, (b) Full gradient norm of empirical loss for training, (c) Empirical loss value for training, and (d) Accuracy score for test to train ResNet-18 on the Tiny ImageNet dataset.

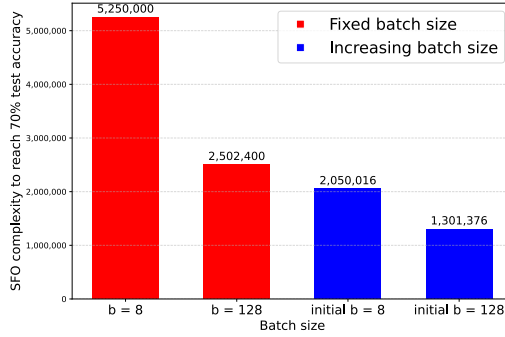
A.5. Visualization of SFO complexity (Tables 2-4)



(a) SFO complexity (Table 2)



(b) SFO complexity (Table 3)



(c) SFO complexity (Table 4)

Figure 7: (a) SFO complexity to reach gradient norm thresholds with $b=8$ (corresponding to Table 2), (b) SFO complexity to reach gradient norm thresholds with $b=128$ (corresponding to Table 3), (c) SFO complexity to reach 70% test accuracy (corresponding to Table 4). Results are obtained using ResNet-18 on the CIFAR-100 dataset.

A.6. Results on wall-clock time

Since wall-clock time is an important practical metric of optimization efficiency, we additionally measured it under the same experimental settings as in Table 2, using CIFAR-100 with ResNet-18 and NSHB, with the maximum batch size set to 1024.

For the fixed batch size $b = 8$ and the increasing batch size starting from $b = 8$ and doubling every 20 epochs, the results are as follows:

Table 5: Wall-clock time to reach gradient norm thresholds ($b = 8$)

Method	Time to reach $\ \nabla f(\theta_t)\ < 0.1$ (h:mm:ss)	Time to reach $\ \nabla f(\theta_t)\ < 0.05$ (h:mm:ss)
Fixed batch size ($b = 8$)	2:22:58	3:40:05
Increasing batch size (initial $b = 8$)	0:49:50	0:54:03

All experiments were conducted on the same computing server, using identical hardware and software environments to ensure a fair comparison. These improvements in wall-clock time are consistent with the reductions in SFO complexity reported in the main paper.

A.7. Assessing the validity of experimental learning rates in light of Theorem 3.2

The learning rate condition in Theorem 3.2 can be expressed as

$$\eta \leq \max \left\{ \frac{1 - \beta}{2\sqrt{2}\sqrt{\beta + \beta^2 L}}, \frac{(1 - \beta)^2}{(5\beta^2 - 6\beta + 5)L} \right\}.$$

Substituting $\beta = 0.9$ into the first term yields approximately

$$\eta \leq \frac{0.0270}{L}.$$

This upper bound depends on the smoothness constant L , which is typically unknown and difficult to estimate accurately in practice. When L is large, the bound becomes more restrictive, requiring a smaller η . Conversely, when L is small, the condition allows for a relatively larger η , under which our empirical choice of $\eta = 0.1$ still appears to fall within a theoretically reasonable range. Consequently, the learning rate setting adopted in our experiments can be regarded as theoretically justified across a wide range of possible values for L , and it is consistent with common empirical practices. Although the exact satisfaction of the theoretical bound cannot be rigorously verified, the consistent convergence observed in our experiments suggests that our setting lies within a reasonable range.